Modélisation d'une ressource termino-ontologique de domaine pour l'annotation sémantique de tableaux

Patrice Buche*, Juliette Dibie-Barthélemy**
Liliana Ibanescu**, Abir Saïd**

*INRA - UMR IATE, 2, place Pierre Viala,
F-34060 Montpellier Cedex 2, France
LIRMM, CNRS-UM2, F-34392 Montpellier, France
Patrice.Buche@supagro.inra.fr
**INRA - Mét@risk & AgroParisTech, 16 rue Claude Bernard,
F-75231 Paris Cedex 5, France
{Juliette.Dibie,Liliana.Ibanescu}@agroparistech.fr

Résumé. Nous proposons dans cet article une modélisation d'une ressource termino-ontologique (RTO) de domaine, guidée par la tâche d'annotation sémantique de tableaux. L'annotation d'un tableau consiste à annoter ses cellules, pour pouvoir ensuite identifier les concepts représentés par ses colonnes et enfin identifier la ou les relations n-aires qu'il représente. La RTO proposée permet d'une part de modéliser dans sa composante lexicale les termes utilisés pour l'annotation des cellules en intégrant la gestion des synonymes et du multilingue, et, d'autre part, de modéliser dans sa composante conceptuelle les concepts symboliques, les concepts numériques et les relations n-aires, qui sont propres au domaine étudié.

1 Introduction

L'intégration de données permet d'accéder de manière unifiée à des sources multiples, hétérogènes en syntaxe, schéma ou sémantique. Le but de l'intégration de données est de faciliter l'accès et la réutilisation d'un ensemble de sources. La notion centrale sur laquelle reposent les recherches actuelles en intégration sémantique de données est la notion d'ontologie, une ontologie étant une spécification formelle et explicite d'une conceptualisation partagée (Studer et al., 1998). Dans Reymonet et al. (2006), trois paramètres influencent la modélisation d'une ontologie : la tâche à réaliser, le domaine d'intérêt et l'application. Dans notre cas, l'application est la construction d'entrepôts de données thématiques ouverts sur le Web pour l'aide à la décision. Notre domaine d'intérêt est la sécurité alimentaire. Nous aurions pu, pour modéliser notre ontologie, reprendre le thesaurus AGROVOC [1] conçu et maintenu par la FAO et qui définit un vocabulaire contrôlé, multilingue et structuré couvrant tous les domaines de l'agriculture, de la pêche et de l'alimentation. Mais ce thesaurus, à vocation très généraliste, n'est pas assez spécifique pour couvrir le sous-domaine de l'alimentation qui nous intéresse (moins

1. http ://aims.fao.org/website/AGROVOC-Thesaurus/sub

de 20% des termes de notre ontologie sont présents dans AGROVOC). Enfin, la tâche à réaliser est l'extraction et l'annotation sémantique de données du Web, ceci afin d'enrichir des bases de données locales. Nous proposons dans cet article la modélisation d'une Ressource Termino-Ontologique (RTO) guidée plus précisément par la tâche d'annotation sémantique de tableaux du Web. Lorsqu'on annote un tableau, on cherche à reconnaître les concepts représentés par ses colonnes et éventuellement la ou les relations sémantiques existantes entre les concepts ainsi identifiés. La classification des colonnes s'appuie sur le contenu de leurs cellules ; elle consiste à annoter les colonnes soit avec des concepts d'une ontologie existante (Hignette et al., 2009), soit à l'aide de mots clefs (Cafarella et al., 2008) ou encore de méta données (Liu et al., 2007). L'identification des relations présentes dans le tableau consiste soit à reconnaître des relations existantes dans une ontologie (Tenier et al., 2006; Hignette et al., 2009), soit à en extraire de nouvelles (Embley et al., 2002; Pivk et al., 2004). Comme dans Hignette et al. (2009), nous nous intéressons à l'annotation de tableaux guidée par une ontologie de domaine, mais nous proposons une nouvelle structuration de l'ontologie, en RTO, dans laquelle les concepts et le lexique sont bien distingués. Cette RTO est composée d'une part de concepts pour annoter les colonnes et les relations n-aires entre ces colonnes et d'autre part d'un lexique de termes pour annoter le contenu des cellules. Une partie de la RTO est générique et le reste dépend du domaine d'application. Nous avons utilisé une approche *top-down* pour la modélisation de ce domaine : les concepts et les relations génériques ont d'abord été modélisés et ensuite les concepts spécifiques. C'est une approche manuelle qui a conduit à une ontologie de très bonne qualité (Sure et al., 2009). Nous présentons dans le paragraphe 2 la modélisation de la RTO et ensuite dans le paragraphe 3 sa représentation en OWL. Ce travail sera illustré par un exemple tiré du domaine d'application concernant le risque chimique dans les aliments.

2 Modélisation de la RTO

La modélisation de la RTO proposée dans cet article est guidée par la tâche d'annotation sémantique de tableaux dans un domaine donné. Un tableau est composé de colonnes, elles-mêmes composées de cellules. Les cellules d'un tableau peuvent contenir i) des termes désignant des concepts symboliques du domaine, ou ii) des termes désignant des concepts numériques du domaine : ce sont des valeurs numériques, souvent suivies de termes désignant une unité de mesure. L'annotation sémantique d'un tableau consiste à annoter le contenu de ses cellules, pour pouvoir ensuite identifier les concepts symboliques ou numériques représentés par ses colonnes et enfin identifier la ou les relations sémantiques n-aires qui existent entre ses colonnes. Une ressource termino-ontologique (RTO) est une ressource comportant une composante conceptuelle, l'ontologie, et une composante lexicale, la terminologie. Chaque composante comporte une partie générique et une partie spécifique qui dépend du domaine d'application. La figure 1 présente la partie générique de la RTO. Nous présentons dans la suite la composante conceptuelle de la RTO puis sa composante lexicale.

2.1 La composante conceptuelle de la RTO

La partie générique de la composante conceptuelle (voir la figure 1) permet de classer les concepts en trois catégories : les concepts simples, les concepts unités et les concepts composés. Les concepts simples regroupent les concepts symboliques et les concepts numériques. La

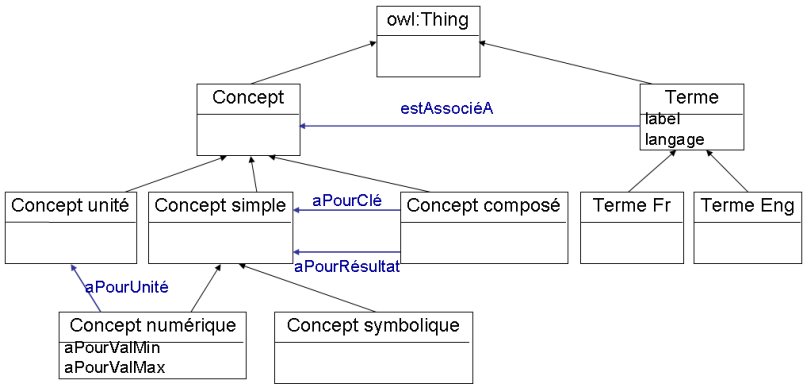

FIG. 1 – *Représentation graphique de la partie générique de la RTO.*

partie spécifique de la RTO permet de définir l'ensemble des concepts spécifiques au domaine d'application étudié. Ils apparaissent dans la RTO en tant que sous-concepts des concepts de la partie générique.

Un concept unité permet de représenter la signification d'une unité de mesure. La classification des concepts unités suit celle du Système international d'unités [2], en unités de base et unités dérivées.

Exemple 1 *Pour le domaine du risque chimique dans les aliments* ng *et* μg *sont des sous-concepts du concept* unités de base *et* ng/g, $\mu g/g$ *et* $\mu g/l$ *sont des sous-concepts du concept* unités dérivés.

Un concept numérique permet de représenter la signification d'une valeur numérique. Un concept numérique peut être exprimé à l'aide d'un ensemble d'unités de mesure et d'un intervalle de valeurs possibles.

Exemple 2 *La figure 2 présente un sous-ensemble des concepts numériques de la RTO correspondant au domaine du risque chimique. Les concepts spécifiques au domaine apparaissent dans les feuilles de la hiérarchie de concepts. Par exemple, le concept numérique* Contamination level *a pour intervalle de valeurs possibles* $[0, 10]$. *Il est de plus associé aux concepts unités dérivés* ng/g, $\mu g/g$, $\mu g/kg$, $\mu g/l$ *(non représenté dans la figure).*

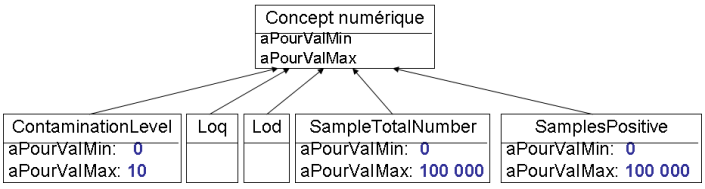

FIG. 2 – *Les concepts numériques dans le domaine du risque chimique.*

2. http ://www.bipm.org/fr/si/

Un concept symbolique permet de représenter la signification d'un terme. Les concepts symboliques sont reliés entre eux par la relation *sorte de*.

Exemple 3 *La figure 3 présente un extrait de la hiérarchie des concepts symboliques dans le domaine du risque chimique. Les concepts spécifiques au domaine sont les sous-concepts de* Concept symbolique*. Par exemple,* Food Product *et* Cereal *sont deux concepts symboliques spécifiques,* Cereal *étant une* sorte de Food Product*.*

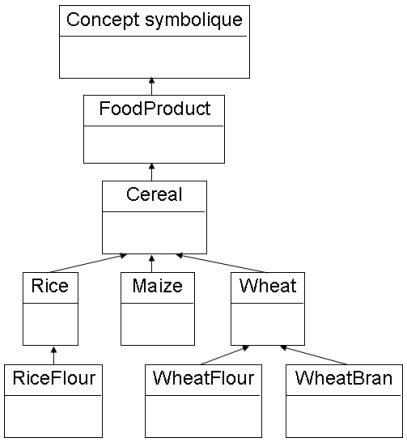

FIG. 3 – *Un extrait de la hiérarchie de concepts symboliques dans le domaine du risque chimique.*

Un concept composé permet de représenter une relation n-aire entre des concepts simples. Un concept composé est défini par son nom et sa signature. La signature est définie par un domaine et un co-domaine : le domaine est composé d'un ou plusieurs concepts simples, appelés les concepts d'accès, tandis que le co-domaine n'est composé que d'un seul concept simple, appelé concept résultat. La limitation du co-domaine à un seul concept simple se justifie par le fait qu'un concept composé, représentée dans un tableau, caractérise en principe un lien sémantique entre des concepts simples avec un seul résultat, comme par exemple un résultat d'expérimentations avec plusieurs facteurs d'entrées. Si un tableau contient plusieurs colonnes résultats, alors il est représenté par autant de concepts composés qu'il contient de résultats.

Exemple 4 *Le concept composé* Contamination Range *est un concept spécifique, sous-concept du concept générique* Concept composé*. Il a pour domaine les concepts symboliques* Food Product *et* Contaminant *et le concept numérique* Samples total number*, et, pour co-domaine, le concept numérique* Contamination level*. Il représente le niveau moyen de contamination d'un produit alimentaire par un contaminant pour un nombre d'échantillons donné.*

2.2 La composante lexicale de la RTO

La composante lexicale de la RTO contient l'ensemble des termes du domaine d'application. Ces termes sont regroupés selon leur langue d'origine (par exemple le français ou l'anglais). On définit un terme comme un ensemble de mots qui dénotent un concept. Un terme

peut désigner un concept simple, un concept unité ou un concept composé. Un terme doit être associé à au moins un concept ; il peut être associé à plusieurs concepts.

Exemple 5 *Le terme anglais "Food product", le terme français "produit alimentaire" et le terme français "aliment" dénotent le concept symbolique* Food Product*.*

3 La RTO en OWL

La traduction en OWL-DL [3] de la RTO, dont la partie générique est présentée dans la figure 1, est obtenue de la manière suivante.

Les concepts apparaissant dans les rectangles sont traduits par des classes OWL. Les flèches, non étiquetées, représentent les relations de spécialisation entre classes. Nous utilisons également l'opérateur de disjonction pour distinguer les classes deux par deux. Les classes *Concept* et *Terme* sont des sous-classes disjointes de la classe *owl :Thing*. Les classes *Concept unité*, *Concept simple* et *Concept composé* sont des sous-classes disjointes de la classe *Concept*. Les classes *Concept numérique* et *Concept symbolique* sont des sous-classes disjointes de la classe *Concept simple*. Les classes *Terme Fr* et *Terme Eng* sont des sous-classes disjointes de la classe *Terme*.

Les flèches étiquetées avec le nom de la propriété représentent les propriétés objet (*owl :ObjectProperty*) entre classes. La relation de dénotation entre termes et concepts de la RTO est représentée par la propriété objet *estAssociéeA* qui a pour domaine la classe *Terme* et pour co-domaine la classe *Concept*. Tel que recommandé par Noy et al. (2006), la signature d'un concept composée, qui caractérise une relation n-aire entre concepts simples, est représentée par les propriétés objet *APourClé* et *APourRésultat*. La propriété objet *APourClé* permet d'associer un concept composé à chacun des concepts simples de son domaine. La propriété objet *APourRésultat* permet d'associer un concept composé à l'unique concept simple de son co-domaine. La propriété objet *APourUnité* permet de représenter les concepts unités qui sont associées à un concept numérique. Les propriétés typées (*owl :DatatypeProperty*) *aPourValMin* et *aPourValMax* permettent d'associer à un concept numérique son intervalle de valeurs possibles. Les propriétés typées *label* et *langage* permettent d'associer à un terme respectivement sa chaîne de caractères et la langue dans laquelle il est exprimé.

4 Conclusion

Nous avons proposé dans cet article une nouvelle modélisation d'une RTO guidée par la tâche d'annotation sémantique de tableaux du Web. La modélisation de la RTO comporte une partie générique et une partie spécifique au domaine d'application. Il est donc possible de réutiliser cette modélisation pour des domaines d'applications variés en redéfinissant la partie spécifique du domaine. La modélisation de la RTO proposée permet de bien distinguer les concepts de la terminologie. L'étape suivante de notre travail consistera à valider cette modélisation par rapport à la tâche d'interrogation de l'entrepôt de données thématique ouvert sur le Web dans un contexte multilingue. Dans un avenir proche, nous utiliserons également cette RTO pour extraire des instances de concepts composés de portions de textes contenus dans des

3. http ://www.w3.org/TR/owl-guide/

documents issus du Web. Nous étudierons enfin comment faire évoluer cette RTO d'une part à l'aide d'autres RTO existantes sur le Web et, d'autre part, en fonction de nouveaux besoins identifiés lors de l'annotation, ceci afin de l'améliorer (ajout de nouveaux termes, de nouveaux concepts, modification de la taxonomie, ...).

Références

Cafarella, M. J., A. Y. Halevy, D. Z. Wang, E. Wu, et Y. Zhang (2008). WebTables : Exploring the Power of Tables on the Web. *PVLDB 1*(1), 538–549.

Embley, D. W., C. Tao, et S. W. Liddle (2002). Automatically Extracting Ontologically Specified Data from HTML Tables of Unknown Structure. In *ER 2002*, pp. 322–337.

Hignette, G., P. Buche, J. Dibie-Barthélemy, et O. Haemmerlé (2009). Fuzzy Annotation of Web Data Tables Driven by a Domain Ontology. In *ESWC*, pp. 638–653.

Liu, Y., K. Bai, P. Mitra, et C. L. Giles (2007). TableSeer : Automatic Table Metadata Extraction and Searching in Digital Libraries. In *JCDL '07*, pp. 91–100. ACM.

Noy, N., A. Rector, P. Hayes, et C. Welty (2006). Defining n-ary relations on the semantic web. W3C working group note. http ://www.w3.org/TR/swbp-n-aryRelations.

Pivk, A., P. Cimiano, et Y. Sure (2004). From Tables to Frames. In *ISWC*, pp. 116–181.

Reymonet, A., N. Aussenac-Gilles, et J. Thomas (2006). Tâche, domaine et application : influences sur le processus de modélisation de connaissances. In *Actes d'IC*, pp. 71–80.

Studer, R., V. R. Benjamins, et D. Fensel (1998). Knowledge engineering : Principles and methods. *Data Knowl. Eng. 25*(1-2), 161–197.

Sure, Y., S. Staab, et R. Studer (2009). Ontology Engineering Methodology. In *Handbook on Ontologies*, pp. 135–152.

Tenier, S., Y. Toussaint, A. Napoli, et X. Polanco (2006). Instantiation of Relations for Semantic Annotation. In *Int. Conf. on Web Intelligence*, pp. 463–472.

Summary

We propose in this paper a model of a termino-ontological resource (TOR) in order to achieve the semantic annotation task of Web data tables in a given domain of application. Table annotation consists in annotating its cells in order to identify concepts represented by its columns and finally identify n-ary relations between concepts which are represented by this table. The proposed TOR allows one: (1) to model, in its lexical component, terms used to annotate cells by considering synonyms and multilingual; (2) to model, in its conceptual component, symbolic concepts, numerical concepts and n-ary relations between them.